Evaluating Datasets for ML: Toolkit

Overview of Research

Background and Motivation: The usage of artificial intelligence has increased exponentially with applications in predicting outcomes related to education, employment, housing, and many more social, economic, and financial aspects of our lives. Archival studies have long dealt with large amounts of data and concerns of representativeness, ethics, integrity, and more with the use of data curation methods, theories, and frameworks. Machine learning research (MLR) has pinpointed the data underlying predictive models to be the largest contributor in introducing bias [27, 30, 31]. Emerging studies have advocated for the prioritization of rigorous data curation practices often referred to as “data work” or “dataset development” in MLR [1, 12, 17]. Introducing data curation concepts and principles can therefore improve the transparency and accountability of the dataset creation process within MLR.

Objectives: We assess ML dataset development processes using principles and methods from archival studies and digital curation. We perform a synthesis and organization of existing work to enable the coherent usage of data curation frameworks, a taxonomy of data curation terms used within machine learning research, and a review of gaps and opportunities for data curation in machine learning.

Method: Our research design for this study consists of the following:

1. Systematic concept mapping between data curation and machine learning dataset creation workflows, concepts, and practices,
2. Comparison of gaps and overlaps across these domains,
3. Iterative development of a rubric as a resource to evaluate and reflect on the ML dataset creation process,
4. Application of the rubric to existing datasets, and
5. Evaluation and refinement of the rubric over multiple iterations

Expected results: This project will deepen the scholarly and practical connections between the data curation and machine learning research communities and initiate directions for improvement within MLR's data practices. By systematically evaluating ML data practices through an explicit digital curation lens, we expect to produce three key results:

1. Evaluation results of NeurIPS benchmark datasets will provide a critical assessment of the state of art and its gaps, including a review of changes in curation practices since the establishment of this track in 2021.
2. The evaluation method and rubric, refined through its substantial application in ML, results in a practical toolkit to be shared widely, including a stepwise guidance tool to support better data practices.
3. The rigorous evaluation of data practices will facilitate critical reflections on the norms currently accepted and widely instantiated.

Prospective Contributions: The outcomes from our project present a novel perspective on improving critical documentation practices in machine learning through data curation. Through this project, we aim to further establish the connections between the data curation and machine learning research communities by identifying opportunities for improvement within MLR's data work processes. This leads to larger ethical impact wherein the uptake of the proposed rubric can aid in reducing bias and increasing accountability and transparency in the dataset creation process.

Application Guidance

Scope of application: The rubric is intended for two types of users.

1. Firstly, dataset creators can use the rubric as a resource to prompt and facilitate critical engagement and reflection throughout their dataset creation process.
2. Secondly, existing datasets can be evaluated prior to publishing or reuse by applying the rubric to determine gaps that require further documentation and areas where bias can be introduced. In both cases, we aim for the rubric to be a practical and useful resource for researchers to engage with the dataset creation process using a data curation lens. The rubric was developed for the evaluation of ML datasets and has elements specific to the domain, including: requirements, data annotation, documentation debt, environmental footprint, and structured documentation.

Applying the rubric to your own dataset

The overall process for using the rubric is as follows:

0. Read the rubric to get familiarized with the elements and details that will be needed.
1. Review each element in the rubric individually.
    a. For each element, first assess whether the minimum standard of documentation has been fulfilled. To do this, provide a pass/fail evaluation, where a pass is granted if all aspects

specified under the minimum standard were discussed and a fail if they were only partially discussed or not discussed at all.

   b. Next, assess whether the documentation meets a standard of excellence, only if the minimum criteria received a pass. The standard of excellence is a full/partial/none evaluation. A full is granted if all aspects specified in the standard of excellence column were discussed, a partial is granted if one or more (but not all) were discussed, and a fail if none were discussed.
   c. It is important to note both for points 2a and 2b that the quality of the responses/documentation is not being assessed but rather if the element was considered and reflected on in any capacity. The purpose of the rubric is to demonstrate the dataset creators' thought process and provide transparency so that its reuse is based on a complete understanding of the dataset.
2. For each element, along with the grade, a comment on what specific information was used to determine that grade must be provided. Other comments and questions can also be included.

The evaluation of each dataset can take 30-60 minutes.

Applying the rubric to existing datasets through publications

The overall process for using the rubric is as follows:

0. Read the rubric to get familiarized with the elements and details that will be needed.
1. Gather and review all pertinent information that can be found about the dataset. This will include the research paper, appendices, the linked dataset, and any documentation associated with the externally linked dataset (e.g., README on github).
2. Review each element in the rubric individually by looking for it across all the information gathered in step 1. Some of the elements will be easier to locate than others because they will be titled specifically, whereas others may be discussed at any point.
   a. For each element, first assess whether the minimum standard of documentation has been fulfilled. To do this, provide a pass/fail evaluation, grant a pass if all aspects specified under the minimum standard were discussed and a fail if they were only partially discussed or not discussed at all.
   b. Next, assess whether the documentation meets a standard of excellence, only if the minimum criteria received a pass. The standard of excellence is a full/partial/none evaluation. A full is granted if all aspects specified in the standard of excellence column were discussed, a partial is granted if one or more (but not all) were discussed, and a fail if none were discussed.
   c. It is important to note both for points 2a and 2b that the quality of the responses/documentation is not being assessed nor the correctness of the technicalities but rather if the element was considered and reflected on in any capacity. The purpose of the rubric is to demonstrate the dataset creators' thought process and provide transparency so that its reuse is based on a complete understanding of the dataset and how it was developed.
3. For each element, along with the grade, a comment on what specific information was used to determine that grade must be provided. Other comments and questions can also be included.
4. For each dataset, evaluators must provide a reflection on their overall assessment of the documentation and rigour demonstrated in the dataset creation process.
5. For each dataset, evaluators must provide a confidence rating for their evaluation.

We estimate the evaluation of each dataset will take about 30-60 minutes once you are familiar with the framework.

How to interpret authenticity, reliability, and representativeness

It may be worth noting that the archival and digital curation perspectives that inform the evaluation framework are particularly important to interpreting the meaning of certain dimensions. Above all, the cluster of authenticity, integrity and reliability needs to be understood from this angle. They are closely related aspects, often treated or addressed by similar mechanisms, but they can be seen as analytically separate concepts. Here is an example.

When you download a data set of weather observations from a platform, you may want to verify if the file you have downloaded in fact is the data set you wanted to get, i.e., is it an authentic copy? You may be able to verify this with various checksums, both on the level of the file (e.g. a hashcode of the file, as commonly provided for downloads) and on the level of observations in some cases. In this case, you are concerned with **authenticity** - you want to verify that the data set is what it purports to be.

Authenticity does not guarantee you, however, that the observations in the data set are any good. A good observation of weather data is one that you can rely on to accurately represent how the weather actually was at the temporal and spatial locations covered by the data. Other aspects of goodness are reflected in the many quality standards for data, but when you want the data set to be able to stand in for the facts it represents, you are concerned with **reliability.** In other words, reliability is very much about the relationship of the data to whatever it represents. If the data set is a compilation of social media posts, then reliability will relate to the question whether these contributions were really posted, etc.

Integrity on the other hand refers to questions of tampering, errors, etc. For example, a dataset that lacks integrity is one for which we can not assert that it contains *all* the items it originally contained, or that none of the items have been altered, falsified, or faked.

Consider a textbook case for records and archives for the difference between the three. A *passport* is a document that comes with very special features to prove that it can *stand in for the fact* that you are a citizen of the issuing country. Its **integrity** refers to the question whether it has been tampered with **-** has the photo been peeled off, have pages been removed or added? etc. The passport comes with features to prevent and check integrity. Its **authenticity** refers to the fact that it is indeed a passport of that country and that it indeed asserts the facts it states. Most of its special features are designed to make it easy to verify that (cf. banknotes). But imagine: a government could issue a perfectly authentic passport for a person who doesn’t exist. That would be authentic, but it would not be reliable. The **reliability** rests on the relationship to the person it represents. We trust an authentic passport to be reliable because we trust the processes that governments have instituted and honed over the centuries to ensure that passports are only *issued to* authenticated citizens. But border control will use a machine readable passport to look up and compare the information shown with the information stored in a database. When they do that, they verify reliability. For a deep dive into the archival perspective on what makes records authentic and reliable, see [5, 9].

Consider next a digital photograph taken during sunlight with a pro-grade digital lens reflex camera of a pantone color set of *whites* with standardized, specified colors, where the white balance is erroneously set at ‘fluorescent light’. White balance relates to the color temperature of light: our eyes automatically adjust to different color temperatures, but a digital sensor does not. How an image looks on a screen is the result of computing it. In this case, the colors will not look very white on the photo without corrections to where the ‘white point’ should be located. The photo as taken is an **authentic** photo providing an **unreliable** representation of its subject. If you transfer the photograph yourself out of the camera you can also put in place mechanisms to verify integrity (including fixity checks and integrity checks using hash sums and the like on the file).

If you notice the error in color and then manually edit the binary code of the RAW file to set the white balance to the correct ‘sunlight’ setting, the photograph would in fact *lose* the property of ‘integrity’ since it has been tampered with (the hashcodes won’t match), and it would *lose* the property of ‘authenticity’ since that was not the original setting, but it would gain in ‘reliability’ since the resulting color rendering would be a more accurate representation of how the colors should look. In this particular case, the fact that the subject of the photograph is standardized provides a *ground truth* that aids in verifying and assessing the photograph. Professional image processing software will be able to document both the ‘as-taken’ setting and ‘to-use’ setting of the photograph. Most photos, of course, are of subjects where this is much harder, and if the photograph is directly processed into a JPEG file, correcting white balance is much more difficult.

Finally, **representativeness** is related to reliability but its perspective is much more narrowly focused on the question whether a data set accurately *represents* the overall set of observations or entities that it claims to be a sample of. For instance, for a data set of social media posts, the question will arise if it’s representative of all platforms, all users, all topics, all media types, or various combinations of dimensions. All the statistical concepts around sampling apply as usual. Other data sets are not sampled out of an identified population but claim to stand for a general category so that representativeness is evaluated analytically, and so on.

How to interpret findability, accessibility, interoperability, and reusability (FAIR)

Note that this group of criteria are a direct representation of the widely used FAIR principles [33] for research data sets, adopted and adapted for machine learning. We provide a simple checklist to assess whether the documentation of the dataset discusses the application of FAIR principles. This checklist is derived from the following tools and resources:

- Minglu Wang and Dany Savard. 2023. The FAIR Principles and Research Data Management. (September 2023). https://doi.org/10.5206/EXFO3999
- FAIR data maturity model
- https://zenodo.org/records/5111307#.Yj3Vi5rMI-Q
- https://ardc.edu.au/resource/fair-data-self-assessment-tool/
- https://fairaware.dans.knaw.nl/

1. Findable
    a. A globally unique (cannot be reused by someone else) and persistent (valid over time) ID (like DOI) is assigned to the data.
    b. The dataset is described by metadata (PID, license, description, provenance, etc.). Further guidelines and definitions of provenance can be found from the DCMI and our glossary.
    c. The metadata specifies the identifier.
    d. The metadata and data is stored in a searchable repository.
2. Accessible
    a. The identifier navigates to the metadata and data.
    b. Retrieval of the data is specified by a standard communications protocol (i.e., all information and tools that are required are communicated to access the content of the dataset) which is open and free to access.

    c. The communications protocol specifies the authentication and authorization procedure, if needed (i.e., if the dataset is not open and free-to-access, the protocol specifies how access would be granted).
    d. The metadata record is available even if the data is not.
3. Interoperable
    a. Metadata and data are *in principle* readable by humans and machines (i.e., has a structured format, open standard).
    b. Metadata and data use controlled vocabularies (standardized and universal terms for indexing and information retrieval). Metadata standards can be found in the RDA Metadata Standards Catalog (https://rdamsc.bath.ac.uk/).
    c. Metadata and data is linked to other metadata and data using qualified references (i.e., relationship to the resource is specified).
4. Reusable
    a. Metadata and data are well-described as per domain-relevant standards, have detailed provenance (where did the data come from, who collected it, when, etc.), and clear and accessible license and usage information.

Guiding Principles

We specify the following principles as "rules of thumb" to guide the evaluation of datasets:

1. Evaluate explicit documentation

Evaluations should be made on the basis of documentation provided by the dataset creators, rather than performing evaluations ourselves.

2. Provide traceable comments.

The comments provided in the rubric to support the grade for each element should make recoverable the basis for the evaluation.

3. Minimum is easy, excellence is hard.

The evaluations for the minimum standard are meant to be *generous*. The evaluation should consider any amount of documentation as a sufficient indicator of reflection for that element. Therefore, meeting the minimum standard should be relatively easy. On the other hand, the standard of excellence criteria advocates for a high level of criticality, which is significantly harder to attain (compared to the minimum standard). The evaluations should therefore only grant a 'Full' if all criteria are satisfied.

4. Don't make excuses.

If there is no documentation provided to evaluate an element, then don't make excuses for the dataset creators and evaluate it yourself or think of it as unnecessary. If you truly feel the element does not apply for that dataset, then that means it's feedback for the rubric and that the element needs further work so it applies to all types of datasets.

Reflections & recommendations

In addition to the instructions on the process of using the rubric to evaluate datasets, the following recommendations are provided based on common reflections, challenges, and questions:

1. Completing an evaluation using the rubric requires iteration. A single pass through the rubric is often insufficient especially for datasets that include various sources of documentation. The first iteration should be a step-by-step completion of each element in the rubric by looking for relevant

information, keywords in the research paper or other dataset documentation. However, in doing so, sections of the documentation may be missed. It is therefore suggested to first evaluate the dataset by applying the rubric sequentially and then reviewing all the dataset documentation sequentially. The final step should be iterating as needed and zooming out.

2. The evaluation of elements will be interconnected, there can be notes to refer to the comment for another element.
3. If a context document is provided, it must be used to evaluate the elements. Although, the document will only provide information to fill in gaps rather than be sufficient to completely evaluate any element.
4. None of the elements should receive an N/A comment or grade.
5. The standard of excellence criteria should only be evaluated if the minimum standard criteria passes.
6. A failure for any element should be not provided based on the quality of the dataset but rather the documentation and reflection on the process of developing the dataset. For example, if the documentation acknowledges that the sample is not representative and can therefore introduce a bias- this is not considered a ‘Fail’.
7. It is important to not evaluate the technical details provided but only evaluate the documentation. This means that evaluators should refrain from inferring the thought process or intention of the dataset creators based on their technical understanding of why the creators would develop their dataset in one way versus another. It is key to rely on the explicit documentation only. This is important because the rubric assesses critical reflection around the dataset process not the quality of the dataset developed.

FAQ

1. Is there a difference between labeling and annotation?

Please refer to the glossary for definitions differentiating the two terms. The rubric doesn’t require evaluation of the “labeling” process if the dataset does not have labels.

2. How to evaluate consistency and timeliness for suitability?

Data quality is often defined as fitness for purpose and is multi-dimensional, meaning that it’s measured through more than one data quality dimension such as accuracy, completeness, etc. Suitability, in the rubric, evaluates whether dataset creators ensure that their dataset’s quality meets the purpose defined. For example, a dataset of math problems may not require timely data but may require consistent data (i.e., data presented in the same format). For standard of excellence, multiple data quality dimensions will apply for evaluation but potentially not all.

3. Is representativeness applicable to synthetic data?

Representativeness is still applicable to synthetic datasets because synthetic data is still representative of reality. However, this is a *conceptual* representativeness rather than a *statistical* one.

4. Why does the evaluation criteria for authenticity discuss data processing specifically?

Data processing alters the authenticity of a digital object. Authenticity is dependent on the bits of information in a file. For example, if you download a dataset with a hash code and make copies of it, all copies will have the same hash code. However, if you perform data processing (which changes the bits), the hash code will no longer be the same. In the rubric, for the minimum standard, you evaluate whether the dataset creators validate and verify the authenticity of the data they are collecting. Whereas for standard of excellence, you evaluate whether they have processes to ensure people that reuse their dataset are able to claim authenticity (i.e., maintaining the chain of authenticity).

5. For the data quality elements, are we evaluating that the dataset is suitable, authentic, has integrity, is representative, and is reliable OR that the dataset creators discuss their processes for ensuring these? If there is no mention of these qualities specifically, how do we evaluate them?

For data quality elements, you are evaluating whether the dataset creators discussed their processes for ensuring that their dataset is suitable, authentic, reliable, has integrity, and the extent to which it is representative (and why if it is not). Remember the guiding principle- “evaluate explicit documentation”. We have added another guiding principle- “don’t make excuses”. If no documentation is provided for these data quality elements, then don’t make excuses for the dataset creators and evaluate it yourself or think of it as unnecessary. If you truly feel the element does not apply for that dataset, then that means it’s feedback for the rubric and that the element needs further work so it applies to all types of datasets.

6. Does hosting a dataset on huggingface make it ‘findable’?

It depends, if it’s hosted on huggingface but does not have a persistent identifier like a DOI, then it is not findable. See next question.

7. Why are URLs not acceptable for findability?

URLs are not considered “findable” because of the high likelihood of link rot (that the link over time will no longer be available). There are studies that show that academic papers are highly perceptible to link rot, eg: see [18]. Instead, we want persistent identifiers like DOIs to make sure the dataset is findable in the future.

8. What is the difference between findability and accessibility?

Findability is about a dataset being easily located. For example, if a publication provides a zenodo link to a dataset, that would make it findable (zenodo assigns a DOI to everything it publishes). So here we’re looking for a dataset being easily located, indexed, catalogued, etc.

Accessibility is about whether a dataset can be opened and used and read. For example, is it in a format you can read, can you download it (i.e., is it retrievable), is the access blocked off via password-protection, are there access and authorization protocols?

A dataset would then be findable if there was a link pointing to it but not accessible if you couldn’t open it because you didn’t have the password for it and there was no documentation of an access protocol. On the other hand, if a dataset was open-access (eg, through github) but didn’t have a persistent identifier (eg DOI) and wasn’t indexed in a repository like zenodo then it would be accessible but not findable. Since accessibility rests on *accessing* the content, a URL alone is not enough to make it accessible either. So even if the dataset is available through github there must be other documentation that provides any further information needed to access the content and metadata.

9. Can you provide further clarification for evaluating interoperability (especially standard of excellence)?

For the minimum standard, the documentation must explain how the dataset can be integrated with other data and workflows. An example of that is that the data can be exported to popular, standard formats. For the standard of excellence, the data and metadata must use controlled vocabularies and link to other resources with qualified references. For example, metadata can be created using controlled vocabularies like the W3C's Data Catalog Vocabulaire (DCAT) model which defines terms like dataset vs data service, catalog (as a subclass of dataset), and so on. Please see this blurb from FAIR about qualified references:

> “A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data, balanced against the time/energy involved in making a good data model. To be more concrete, you should specify if one dataset builds on another data

> set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly cited (i.e., including their globally unique and persistent identifiers).” [11]

Zenodo also has a webpage that describes how it fulfills the FAIR principles for its datasets [34].

Rubric

	CURATORIAL ELEMENT	DESCRIPTION	DOCUMENTATION LEVEL	
			Criteria to meet minimum standard	Criteria to meet standard of excellence
SCOPE				
1	Context, purpose, motivation	This information explains the purpose of dataset creation for the specified domain.	Documentation discusses the problem domain, what problems the new dataset addresses, the relevance of those problems, and the need for a new dataset in comparison to existing datasets.	Documentation explains how the context of the dataset affects possible reuse and includes reflection on the dataset creators' awareness of social, political, and historical context.
2	Requirements	The translation process from a "real-world" problem to a "ML problem" for which the dataset is created [24, 26] consists of numerous decisions, expertise, and worldviews that should be documented in order to understand the context in which the problem situation was framed.	Documentation states how the problem was formulated and how the dataset creation plan was generated.	Documentation includes reflection on how the problem formulation introduces intrinsic biases.
ETHICALITY AND REFLEXIVITY				
3	Ethicality	Ethical considerations are critical to the fair and accountable creation and (re)use of datasets.	Documentation discusses how the benefits of creating the dataset outweigh any harms of creating it (see proportionality principle), and it discusses informed consent if the dataset is about humans.	Documentation goes beyond requirements listed in ethics framings like guidelines/policies/checklists. For example, documentation discusses alternate methods of dataset creation that were not used because of potential ethical harm.
4	Domain knowledge & data practices	Creating a dataset involves, often tacit, expertise about one or more domains as well as data practices. Articulating both types of nuance required in dataset development makes data work more transparent [12, 15, 24, 28, 32].	Documentation states the domain-specific expertise and data skills required in developing the dataset.	Documentation discusses the required expertise needed to understand the intended purpose of the dataset and to reuse it.
5	Context awareness	Context awareness demonstrates an understanding of the subjective, non-neutral nature, and situatedness of data.	Documentation includes a positionality statement.	Documentation adopts a reflexive approach to dataset development. For example, documentation discusses how field epistemologies impact assumptions, methods, or framings.
6	Environmental footprint	This element is for dataset creators to reflect and quantify the footprint of their dataset creation process [1].	Documentation contains a quantitative assessment of environmental footprint and clearly defined scope of what was measured.	Documentation includes a lifecycle assessment and the corresponding environmental footprint, and an assessment of design choices and rationale for the choices.
DATA PIPELINE				

7	Data collection	Disclosing data sources is essential in the data collection process. Further reflection on the process of selecting those sources can reveal important interpretive assumptions [24] and historical and representational biases [15].	If data was collected, documentation states how and why data and metadata were collected from the data source(s). If data was synthesized, documentation discusses: 1) how and why the data was synthesized and 2) whether the data was synthesized to match labels, if used.	If data was collected, documentation discusses the process of defining criteria for selecting data source(s), specifies the criteria, explains why those criteria were chosen, and how the selected data sources are evaluated against these criteria. If data was synthesized, documentation includes a reflection on potential intrinsic biases of the synthesis process, how the synthesis process shaped the features of the data, the limitations of the synthesis process, and how the synthesized data relates to the real-world distribution of the data it represents.
8	Data processing	Data processing involves cleaning, transforming, and wrangling data. Data processing decisions have impacts on the ultimate “cleaned” data that is used [21, 24]. Detailed documentation of this process enables outcomes of the model to be traced back to processing decisions.	Documentation discusses the process of cleaning, transforming, or wrangling data.	Documentation goes beyond what is done to discuss how the decisions about data processing were made and why, and potential impacts of the processing decisions.
9	Data annotation	Data annotation or labelling, regardless of the guidelines provided to reduce worker bias, can lead to disagreements on how data should be annotated (either between annotators or between dataset creators and annotators).The inclusion of this documentation highlights what is considered the “ground truth” [4, 24, 25] by the dataset creators which impacts how annotation is performed [16].	Documentation discusses the process of annotation. If any labels are used, the documentation includes the following: If labels are derived from the data: documentation discusses how data was interpreted to generate labels. If the labels were created first and the data was derived from the labels: documentation discusses how the relationship of the data to the labels was verified. If labels are obtained from elsewhere: documentation discusses where they were obtained from, how they were reused, and how the collected annotations and labels are combined with existing ones.	Documentation discusses the process of annotation with depth and reflexivity by including a reflection on how annotations (including labels, if used) represent differing worldviews and social backgrounds. Additionally, if labels are derived from the data: documentation discusses how the labels are robust, i.e., not sensitive to variability and how disagreements on annotation were reconciled.
DATA QUALITY				
10	Suitability	Suitability is a measure of a dataset’s quality with regards to the purpose defined.	Documentation discusses how the dataset is appropriate for the defined purpose.	Documentation discusses how dimensions such as accuracy, completeness, timeliness, and consistency contribute to the quality of the dataset in being used for the defined purpose. For example, timeliness

				(i.e., age) of data should be appropriate for the defined purpose.
11	Representativeness	Representativeness is a measure of how well a sample set of data represents the entire population. Sampling procedures and decisions about data sources can introduce extrinsic bias [24]. For example, choosing Reddit or Twitter as a data source can perpetuate dominant social biases rather than being a representative sample of the target population [1].	Documentation defines the population and discusses the extent to which the sampling procedure is representative of the population.	Documentation includes reflection on how the dataset creation process overall, and the sampling procedures specifically, affect extrinsic bias.
12	Authenticity	Authenticity of a dataset is about whether the dataset "is what it purports to be" [5, 7, 8, 13, 29], which is a responsibility of dataset creators [20]. Authenticity can be established by assessing the identity and the integrity of the record [5, 6, 10, 14, 19, 22]. Integrity of a dataset is about whether "the material is complete and unaltered" [2, 3, 9, 13, 23].	Documentation discusses how authenticity has been established and maintained, i.e., • Has the identity and origin of all data been verified? • For data that is obtained, it is clear how the dataset creators have verified the identity of the dataset they reuse. • For data that is generated, it is clear how they have been created and by whom. • Has the integrity of all data been verified? • For data that is processed in any way, it is clear how processing steps may have impacted integrity.	Documentation states how others can establish the authenticity of this dataset, i.e., • Documentation provides a persistent identifier and provenance information for the dataset in order for reusers to establish identity. • Documentation provides mechanisms for reusers to verify the integrity of their dataset.
13	Reliability	Reliability is about how well the dataset is "capable of standing for the facts to which it attests" [5], i.e., how certain we can be that its data points reflect what they represent.	Documentation discusses how the reliability of the dataset has been established and maintained, including the verification steps taken to ensure reliability, where necessary, i.e., • It is clear for each data element what synthetic or real-world phenomenon it represents.	Documentation states how others can establish the reliability of the dataset, i.e., • Documentation provides mechanisms to enable verification of what synthetic or real-world phenomenon each data element represents.
14	Structured documentation	Context documents in standardized structures provide information on the content of the dataset which is critical in establishing its usage in a well defined format.	Documentation includes a standardized context document. Acceptable formats include context documents that follow an established structure such as datasheets, data statements, and nutrition labels.	The context document addresses all mandatory items.
DATA MANAGEMENT				

15	Findability	Ensuring findability is about enabling the dataset to be discovered for reuse after its development [33].	Documentation discusses how the dataset is findable by providing a globally unique and persistent identifier (URLs are not persistent).	Documentation includes metadata and both the metadata and data are stored in a searchable repository.
16	Accessibility	Accessibility is about enabling the dataset to be obtained after its development [33].	Documentation states all information and tools required to access the content of the data, and the identifier navigates to the metadata and data.	Documentation includes a communications protocol, an authentication and authorization procedure, and provides metadata that will be available even if data access is removed.
17	Interoperability	Interoperability ensures that the dataset can be integrated with other applications and workflows [33].	Documentation discusses how the dataset integrates with other data, workflows, applications, etc. (i.e., that both the metadata and data are readable by humans and machines).	Documentation has metadata and data that both use controlled vocabularies and link to other resources using qualified references.
18	Reusability	Ensuring reusability requires providing information such as relevant provenance and usage [33].	For both metadata and data, provenance information includes at least all of the following: 1) where the data came from, 2) who collected it, and 3) when it was collected.	Documentation has metadata and data that are both described using domain-relevant standards, state license and usage information, and provide additional provenance documentation as described by FAIR best practices.

Rubric Worksheet

	CURATORIAL ELEMENT	DOCUMENTATION LEVEL			
		Criteria to meet minimum standard		Criteria to meet standard of excellence	
		Pass/Fail	Comments	Full/Partial/None	Comments
SCOPE					
1	Context, purpose, motivation				
2	Requirements				
ETHICALITY AND REFLEXIVITY					
3	Ethicality				
4	Domain knowledge & data practices				
5	Context awareness				
6	Environmental footprint				
DATA PIPELINE					
7	Data collection				
8	Data processing				
9	Data annotation				
DATA QUALITY					
10	Suitability				
11	Representativeness				
12	Authenticity				
13	Reliability				
14	Structured documentation				
DATA MANAGEMENT					
15	Findability				
16	Accessibility				
17	Interoperability				
18	Reusability				

Samples

Please note that the sample evaluations were performed using the version of the rubric at the time of evaluating datasets from round 3. Note also that the description column and cited references are deleted below for space.

Example 1

Paper: FS-Mol: A Few-Shot Learning Dataset of Molecules

	CURATORIAL ELEMENT	DOCUMENTATION LEVEL			
		Criteria to meet minimum standard	PASS/ FAIL	Criteria to meet standard of excellence	Full/ Partial/ None
SCOPE					
1	Context, purpose, motivation	Pass	Paper introduction discusses the problem domain and why a new dataset is needed; see ‘related work’ in paper and appendix B in supplementary material (‘related work details’) for comparison to existing datasets.	Full	Section 7 of paper discusses how dataset can be used outside of its original context (“it is now possible to evaluate… we note that transfer of results to realistic projects is not guaranteed to be successful…”)
2	Requirements	Pass	Section 2 of paper (especially “ 2.2 Desired Attributes of a QSAR Few-Shot Dataset and Benchmark”) explicitly derives design requirements to create the dataset.	Partial	No explicit discussion of intrinsic biases introduced by problem formulation; other approaches to formulating the problem are discussed in ‘related work’ section of paper (discussing other datasets and their features)
ETHICALITY AND REFLEXIVITY					
3	Ethicality	Pass	No discussion of consent (no human data); pg 9 ‘societal impacts’ section discusses benefits of creating the dataset.	Fail	No additional discussion of ethical consideration throughout the paper or supplementary documentation.
4	Domain knowledge & data practices	Pass	On pg 2 of papers, authors state aim to “demonstrate the utility of few-shot learning methods in an important domain, namely QSAR,	Partial	README in GitHub repo discusses activities to be undertaken to re-use the dataset “Hence, in order to be able to run MAT, one has to clone our repository via…” – not directly discussing any domain knowledge needed.

			which does not provide an obvious generic pretraining corpus (such as in NLP or computer vision). The proposed dataset is specifically designed to replicate the challenges of machine learning in the very low data regime of drug-discovery projects" (focus on drug-discovery domain)		
5	Context awareness	Fail	Research goals are described but not positioned relative to researchers' intellectual/political believes; researcher positions not disclosed/no positionality statement included.	None	Failed minimum criteria.
6	Environmental footprint	Fail	No assessment of environmental footprint	None	Failed minimum criteria
DATA PIPELINE					
7	Data collection	Pass	ExtractDataset.ipynb from GitHub repo describes how data were gathered by querying ChEMBL; section 3 of paper explains data acquisition process in detail ("the reason why we remove large assays is…")	Partial	Section B of supplementary material describes other few-shot learning and molecular property datasets (e.g. why they used ChEMBL instead of other sources); no explicit discussion of criteria for source selection, why criteria were chosen, or how other sources were validated against criteria.
8	Data processing	Pass	ExtractDataset.ipynb from GitHub repo describes how data were cleaned and split into test vs validation assays.	Full	Section 3 of paper describes decisions behind data processing (e.g. "In this way, our proposed meta-testing tasks closely mimic the new-lead optimization problem, where a completely unseen task is presented for adaptation.")
9	Data annotation	Pass	"Binary Classification Task" section of paper discusses some annotation activity	None	No discussion of robustness of annotations.
DATA QUALITY					
10	Suitability	Pass	Section 6 and first paragraph of section 7 describe and demonstrate dataset appropriateness for purpose.	Partial	Documentation does not explicitly discuss accuracy/completeness/timeliness of the chosen dataset, but Section 6 of the paper demonstrates the utility of the dataset for its intended purpose by providing "a set of results for all three categories of few-shot learning, with representative methods of the use of this dataset in each".

11	Representativeness	Pass	Section 3 on pg 3 of main paper describes how the ‘sample’ of the dataset is taken from the overall population (the ChEMBL database); also on pg 9 “the few-shot baselines we provide checkpoints and results for are only a representative set, rather than a complete survey of the current state of the field”	None	No explicit discussion of biases.
12	Authenticity	Pass	No explicit discussion of authenticity but extractdataset.ipynb does discuss how initial raw data were obtained (e.g. describes process by which database was queried)	Partial	No explicit discussion of future authenticity/preservation processes, but does discuss in section A of supp material how dataset documentation facilitates re-use more generally.
13	Reliability	Pass	Section 5 of paper discussing benchmarking procedures (i.e. making sure that the dataset is useful for what it’s supposed to be useful for)	Partial	No explicit discussion of reliability management in the context of future re-use; section A of supplementary material discusses how the dataset documentation facilitates re-use.
14	Integrity	Fail	No discussion of dataset integrity or preservation processes (section H of supplementary document does not actually discuss a maintenance plan or means of maintaining accuracy/consistency over time).	None	Failed minimum criteria.
15	Structured documentation	Fail	No standardized context document	None	Failed minimum criteria
DATA MANAGEMENT					
16	Findability	Fail	No persistent identifier provided.	None	Failed minimum criteria
17	Accessibility	Pass	Section F of supplementary material describes computational resources used; GitHub README states the tools and steps required to access data content.	Partial	GitHub repo includes a code of conduct document, as well as protocols for contributing and for security reporting.
18	Interoperability	Pass	README in GitHub repo describes how to use the dataset with “three key few-shot learning methods”; dataset.ipynb describes the machine/human readable metadata.	Full	Dataset.ipynb describes the controlled vocabularies for specific dataclasses (e.g. task_name as a string describing the task each point is taken from)

19	Reusability	Fail	From data contents of GitHub repo it does not appear that data or metadata contain provenance information about where the dataset came from/when/who collected it; license is included in the GitHub repo.	None	Failed minimum criteria.

Example 2

Paper: American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers

	CURATORIAL ELEMENT	DOCUMENTATION LEVEL			
		Criteria to meet minimum standard	PASS/ FAIL	Criteria to meet standard of excellence	Full/ Partial/ None
SCOPE					
1	Context, purpose, motivation	Pass	Paper Introduction and section 6 (Applications) discusses the problems and relevance, and ‘Related Literature’ (section 2) discusses other similar datasets.	Full	‘Applications’ section on pg 6 of supplementary material discusses “multiple applications that can be facilitated by the American Stories dataset”
2	Requirements	Pass	Paper Introduction (pg 2, “To address these limitations, we develop…”) introduces certain requirements.	Partial	On pg 3 of paper ,documentation reflects on the bias potentially introduced by scanning illegible newspapers; other approaches are discussed in Section 2 on Related Literature (but not specifically other approaches the authors considered)
ETHICALITY AND REFLEXIVITY					
3	Ethicality	Pass	Some harms (e.g. offensive language) are discussed in Section 7: Conclusion. Consent is discussed in datasheet (pg 14 of supplementary material)	Partial	Some additional discussion of copyrights/accessibility on pg 3 of paper
4	Domain knowledge & data practices	Pass	Pg. 23 of paper (the datasheet) addresses the professors, research assistants, and students involved in data collection	Partial	Datasheet states “There are a large number of potential uses in the social sciences, digital humanities, and deep learning research”
5	Context awareness	Pass	No positionality statement but several mentions throughout the datasheet showing awareness of social context (“This dataset contains unfiltered content composed by newspaper editors, columnists, and other sources. It reflects their biases and any factual errors that they made.”),	Partial	Section 3 of paper touches on assumptions going into methodological choices (e.g. on pg 3, “We do not OCR ads because…”)

			and section 7 of the paper reflects on the historicity of dataset contents		
6	Environmental footprint	Fail	No environmental assessment.	None	Failed minimum criteria
DATA PIPELINE					
7	Data collection	Pass	Described in ‘Composition’ (pg 11) and ‘Collection Process’ (pg 13) sections of datasheet in supplementary material	Partial	We have a lot of information about how the data were collected, but I still don't see where in the documentation it specifies the criteria they used to select data sources or how data sources were validated against these criteria (e.g. why the library of congress dataset?).
8	Data processing	Pass	Pre-processing section of datasheet (pg 14 of supplementary material) describes process of cleaning and wrangling data	Full	Sections 3, 4, and 5 of main paper discuss the implications of processing decisions (e.g. on computing cost and efficiency)
9	Data annotation	Pass	Student annotation is discussed ins Section 5 ‘Pipeline Evaluation’ of main paper	Full	Student annotations were used as ‘ground truth’ for model training; see pg 5 of supplementary material
DATA QUALITY					
10	Suitability	Pass	Section 5 of paper evaluates the pipeline for accuracy, legibility, and comparison to other OCR engines	Full	See explanation for minimum criteria
11	Representativeness	Pass	Sampling approach discussed in datasheet (pg 13 of supplementary material) – it includes everything in the Chronicling American scan collection.	Full	Section 3 of paper discusses how illegible papers and their inclusion/exclusion in the dataset could bias results.
12	Authenticity	Pass	Pipeline for generating data is included in the Github repo (https://github.com/dell-research-harvard/AmericanStories?tab=readme-ov-file); no explicit discussion of authenticity	None	No explicit discussion of authenticity in future re-use.
13	Reliability	Pass	Section 5 of paper (Pipeline Evaluation) describes verification and validation processes used to ensure reliability.	Full	Maintenance section of datasheet discusses how errors will be corrected in future (and uploaded to HuggingFace)
14	Integrity	Pass	Documentation does not explicitly discuss integrity but datasheet does emphasize that “material is complete and unaltered”	Full	Maintenance section of datasheet describes preservation processes in place (e.g. old versions still accessible via HuggingFace)
15	Structured documentation	Pass	Paper and supplementary material include a datasheet (Gebru et al)	Full	All mandatory components of datasheet are answered.
DATA MANAGEMENT					

16	Findability	Pass	DOI available on HuggingFace page (10.57967/hf/0757)	Full	Data and metadata stored in searchable repo (HuggingFace)
17	Accessibility	Pass	Steps for accessing data listed on HuggingFace page data card and described in ‘Distribution’ section of datasheet (pg 15 of supplementary material	Full	Communications protocol described in ‘Maintenance’ section of datasheet (supp material pg 16)
18	Interoperability	Pass	Pg 4 of paper describes readable formats of metadata and data (“The raw files are in a json format, and the Hugging Face repo comes with a setup script that easily allows people to download both raw and parsed data to facilitate language modeling and computational social science applications.”; lots of metadata info included on HuggingFace page	Full	See HuggingFace page for controlled metadata vocabularies
19	Reusability	Pass	Some provenance information included in metadata (e.g. where it came from, associated newspaper, but not who collected it/when)	Partial	Pg 16 of supplementary material (datasheet) states “The dataset is distributed under a Creative Commons CC-BY license. The terms of this license can be viewed at https://creativecommons.org/licenses/by/2.0/”

Further Readings

The following readings 1) showcase how data curation is discussed in data science and machine learning studies, 2) contain context for relevant data curation terms, concepts, and frameworks, and 3) provide important terminology for ML benchmarks. Readings are listed as required and suggested.

Data Curation in Data Science

A vast amount of literature points to the datasets used for training machine learning models to be the source for introducing bias in model results leading to a call for increased documentation of datasets used in ML. Emerging research has proposed context documents – "interventions designed to accompany a dataset or ML model, allowing builders to communicate with users". The following are types of relevant context documents.

Required:

1. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (November 2021), 86–92. https://doi.org/10.1145/3458723

Datasheets are one of the most popular methods of documenting the process of developing datasets as well as providing a dataset description. This paper is a good introduction to how dataset documentation is evaluated.

Suggested:

2. Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 21, 2020, Honolulu HI USA. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

Madiao et al. developed a resource - checklist for AI fairness - based on findings of current practitioners processes, needs, and requirements for developing fair AI models.

3. Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6, (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

Bender and Friedman develop 'data statements'- a resource for NLP training datasets to be documented in order to mitigate bias and exclusion.

Topics like dataset documentation in ML are often discussed as a part of data practices, data work, or dataset development. The following studies talk about stages of dataset development processes, how data scientists or data workers approach their data work, and the importance and impact of decisions made during the dataset development.

Required:

1. Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–37. https://doi.org/10.1145/3476058

This paper discusses how documentation captures underlying values of data practices in machine learning (specifically computer vision tasks). Specifically, publications are analyzed to understand the documentation and communication of datasets. The findings showcase the practices that are silenced (such as data work, context, positionality, and care) over those that are (wrongly) embraced such as model work, universality, and so on. This reading help reflect on and understand how intrinsic bias can be introduced within datasets.

Suggested:

2. Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021, Yokohama Japan. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

Through interviews with AI practitioners, Sambasivan et al. find that poor data practices in high-stakes AI domains (i.e., practices that do not prioritize data quality) lead to data cascades which are negative impacts of data issues.

3. Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP (January 2022), 1–14. https://doi.org/10.1145/3492853

Miceli et al. discuss that while we often recognize that there is bias in the datasets and their processes used for ML models, it is often ignored that this bias is a result of power inequities. The authors analyze data bias, data work, and data documentation from a "power-aware" framing as compared to a "bias-oriented" one. This paper provides an interesting shift in perspective which further illuminates the importance of reflexivity in data work.

4. Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022, New Orleans LA USA. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644

This paper studies how data processing leads to different types of forgetting and where and how each type of forgetting occurs in the machine learning stack. Forgetting is conceptualized as the practice that occurs when choices are made about what data is kept, what it represents and so forth (therefore by designing a dataset in a given way, we *remember* only its current state, and *forget* the decisions, the erased data, etc.). This is a great paper for a deep dive into the various types of design decisions that impact the eventual dataset.

The previous studies discuss aspects of data curation as dataset development. However, some ML studies have started discussing the importance of data curation by referencing archival studies and digital curation directly. These are included below:

Required:

1. Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In *Proc. of 2021 AAAI/ACM Conf. on AI, Ethics, and Society*, July 21, 2021, Virtual Event USA. ACM, Virtual Event USA, 695–703. Retrieved November 11, 2022 from https://dl.acm.org/doi/10.1145/3461702.3462598

This study discusses principles for ethical data curation based on race critical race theory and data feminism to improve the reflection of power, bias, and values in data processes and thereby improve transparency and accountability of AI systems.

Suggested:

2. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '21*), March 01, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 610–623. . https://doi.org/10.1145/3442188.3445922

This paper discusses the potential risks of language models (and by extension other ML/AI systems). The authors recommend a shift towards careful, reflective practices around datasets and model development along with a greater focus towards documentation.

3. Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 27, 2020, Barcelona Spain. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

This paper highlights that practices from archival studies have experience dealing with consent, power dynamics, transparency, and ethics and that these practices should be adopted into data collection and annotation practices in machine learning.

Data Curation

Data curation involves "maintaining and adding value to digital research data for current and future use". The following studies introduce data/digital curation terminology and the data curation lifecycle model (parallel to ML model pipelines) with the aim to familiarize how the data curation field approaches data work.

Required:

1. Sarah Higgins. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3, 1 (August 2008), 134–140. https://doi.org/10.2218/ijdc.v3i1.48 (skim only)

The paper introduces the curation lifecycle model by emphasizing it as a lifecycle (as opposed to a linear process). Each stage of the model is briefly introduced.

2. Sarah Higgins. 2012. The lifecycle of data management. In *Managing Research Data* (1st ed.), Graham Pryor (ed.). Facet, 17–46. https://doi.org/10.29085/9781856048910.003

This paper discusses each stage in depth including the tasks performed, how each stage leads to the next, and the expected outcomes.

Suggested:

3. Digital Curation Centre. Glossary. *Digital Curation Centre*. Retrieved January 21, 2024 from https://www.dcc.ac.uk/about/digital-curation/glossary

This is a glossary of common digital curation terms - to be returned to as a resource, as needed.

4. Carole L Palmer, Nicholas M Weber, Trevor Muñoz, and Allen H Renear. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data. 16.

This is an introductory paper to the field of data curation and its place within archival studies, library studies, and computer science.

Benchmarking in ML

Benchmarking is often not a well discussed topic in machine learning papers. The below list is compiled to introduce commonly used terms including: benchmark dataset, benchmark tasks, simulator, synthetic dataset, baseline method, benchmark suite, etc.

1. Matthew Stewart. 2023. The Olympics of AI: Benchmarking Machine Learning Systems. *Medium*. Retrieved January 21, 2024 from https://towardsdatascience.com/the-olympics-of-ai-benchmarking-machine-learning-systems-c4b2051fbd2b

Explains terms benchmark, benchmark dataset, benchmark tasks, baseline method, and benchmark suite.

2. Ramona Leenings, Nils R. Winter, Udo Dannlowski, and Tim Hahn. 2022. Recommendations for machine learning benchmarks in neuroimaging. *NeuroImage* 257, (August 2022), 119298. https://doi.org/10.1016/j.neuroimage.2022.119298

Explains benchmark term and concept.

3. Kim Martineau. 2021. What is synthetic data? *IBM Research Blog*. Retrieved January 21, 2024 from https://research.ibm.com/blog/what-is-synthetic-data

Explains term synthetic data.

4. Nataniel Ruiz. 2019. Learning to Simulate. *Medium*. Retrieved January 21, 2024 from https://towardsdatascience.com/learning-to-simulate-c53d8b393a56

Explains term simulator.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 03, 2021. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[2] June M. Besek and Philippa S. Loengard. 2007. Maintaining the Integrity of Digital Archives. *Columbia J. Law Arts* 31, (2007), 267.

[3] Li Cai and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* 14, (May 2015), 2–2. https://doi.org/10.5334/dsj-2015-002

[4] Catherine D'Ignazio and Lauren F. Klein. 2023. *Data Feminism*. MIT Press.

[5] Luciana Duranti. 1995. Reliability and Authenticity: The Concepts and Their Implications. *Archivaria* (May 1995), 5–10.

[6] Luciana Duranti. 1998. *Diplomatics: New Uses for an Old Science*. Scarecrow Press.

[7] Luciana Duranti. 2005. The long-term preservation of accurate and authentic digital data: the INTERPARES project. *Data Sci. J.* 4, (2005), 106–118. https://doi.org/10.2481/dsj.4.106

[8] Luciana Duranti. 2007. The InterPARES 2 Project (2002-2007): An Overview. *Archivaria* (2007),

113–121.

[9] Luciana Duranti and Heather MacNeil. 1996. The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project. *Archivaria* (October 1996), 46–67.

[10] Luciana Duranti and Randy Preston. 2009. International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records. *Rec. Manag. J.* 19, 1 (January 2009). https://doi.org/10.1108/rmj.2009.28119aae.003

[11] GO FAIR. I3: (Meta)data include qualified references to other (meta)data. *FAIR Principles*. Retrieved January 18, 2024 from https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/

[12] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (November 2022), 1–29. https://doi.org/10.1145/3555760

[13] Sarah Higgins. 2009. DCC DIFFUSE Standards Frameworks: A Standards Path through the Curation Lifecycle. *Int. J. Digit. Curation* 4, 2 (October 2009), 60–67. https://doi.org/10.2218/ijdc.v4i2.93

[14] Asen O Ivanov. 2019. The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation: Curation Culture and Evaluative Practice. University of Toronto.

[15] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 27, 2020. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

[16] Julian Posada. 2023. *Platform Authority and Data Quality*. Retrieved from https://www.berggruen.org/ideas/articles/decoding-digital-authoritarianism/

[17] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. (September 2022). https://doi.org/10.2139/ssrn.4217148

[18] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE* 9, 12 (December 2014), e115253. https://doi.org/10.1371/journal.pone.0115253

[19] Brent Lee. 2005. Authenticity, Accuracy and Reliability: Reconciling Arts-related and Archival Literature. (2005).

[20] Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, Wim Hugo, Reyna Jenkyns, Varsha Khodiyar, Maryann E. Martone, Mustapha Mokrane, Vivek Navale, Jonathan Petters, Barbara Sierman, Dina V. Sokolova, Martina Stockhause, and John Westbrook. 2020. The TRUST Principles for digital repositories. *Sci. Data* 7, 1 (May 2020), 144. https://doi.org/10.1038/s41597-020-0486-7

[21] Lydia R. Lucchesi, Petra M. Kuhnert, Jenny L. Davis, and Lexing Xie. 2022. Smallset Timelines: A Visual Representation of Data Preprocessing Decisions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. ACM, Seoul Republic of Korea, 1136–1153. https://doi.org/10.1145/3531146.3533175

[22] H. MacNeil. 2013. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. Springer Science & Business Media.

[23] Reagan Moore. 2008. Towards a Theory of Digital Preservation. *Int. J. Digit. Curation* 3, 1 (August 2008), 63–75. https://doi.org/10.2218/ijdc.v3i1.42

[24] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644

[25] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI*

Conference on Human Factors in Computing Systems, May 06, 2021. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445402

[26] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19*), January 29, 2019. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[27] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (November 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[28] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. 2021. Advances in Neural Information Processing Systems.

[29] Alex H. Poole. 2015. How has your science data grown? Digital curation and the human factor: a critical literature review. *Arch. Sci.* 15, 2 (June 2015), 101–139. https://doi.org/10.1007/s10502-014-9236-y

[30] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

[31] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–37. https://doi.org/10.1145/3476058

[32] Andrea K. Thomer, Dharma Akmon, Jeremy J. York, Allison R. B. Tyler, Faye Polasek, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (November 2022), 414:1-414:29. https://doi.org/10.1145/3555139

[33] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18

[34] Zenodo - Research. Shared. FAIR Principles. Retrieved January 18, 2024 from https://about.zenodo.org/principles/